

京都大學工學研究科
社會基盤工學専攻

2026年度
卒業論文
土木工学専攻
資源工学専攻



Refraction-Aware Gaussian Splatting for Shallow Water Bathymetry from UAV Imagery

京都大学大学院 工学研究科 社会基盤工学専攻
空間情報学講座
宇野 大輝

論文要旨

ACRS か ISPRS の内容を翻訳しているだけ。最後にどこまでできたか踏まえて書く。内容も増えるはず。

水中の3次元地形の計測は水深測量(Bathymetry)と呼ばれている。

浅水域においても、河床・海底の水深測量は、地形変動のモニタリング、ハザードシミュレーション、および水生生物の生息環境評価において極めて重要である。近年、無人航空機(Unmanned Aerial Vehicle: UAV)を用いた写真測量(Photogrammetry)は、広範囲を効率的に調査する手法として注目されている。しかし、空中から水底を撮影する際、水面で発生する光の屈折が、従来の写真測量における幾何学的仮定(共線性条件)を根本的に破綻させるという課題がある。既存の手法は、屈折を完全に正確に物理的正確性を持ってモデル化しない経験的な補正や反復的な後処理に依存するか、あるいは説明可能性を欠くブラックボックス的な深層学習モデルを用いるものが多く、形状の幾何学的忠実性と外観の写実性を両立させることは困難であった。

本論文では、この課題を解決するために、物理的に忠実な二媒質屈折モデルを再構成パイプラインに直接組み込んだ“Refraction-aware 3D Gaussian Splatting”を提案する。本手法の核心的な貢献は、水中の真の位置にある 3D Gaussian を、航空画像上の見かけの位置へと解析的にマッピングする微分可能なパラメータ変換の導入にある。これにより、標準的な 3D Gaussian Splatting の柔軟なフレームワークを維持しつつ、屈折あり画像からの密な 3 次元形状と詳細なテクスチャ情報の復元を実現した。

評価実験では、屈折以外の光学的要因を排除し厳密な検証を行うため、物理ベースのレイトレーシングにより生成された河床のシミュレーションデータセットを用いた。その結果、水深 10 m のスケールにおいて許容誤差 10 cm とした場合の幾何学的 F1 スコアは 96% を達成した。さらに、新規視点合成(Novel View Synthesis)においては、PSNR 25.9 dB、SSIM 0.93 という詳細な屈折なし画像の推定を達成した。提案手法により、平坦な水面条件下において、河川、湖沼、沿岸域の低コストかつ高頻度な 3D モニタリングが可能となり、水域リモートセンシング分野に新たな方法論的基盤を提供する。

目次

TODO

第1章 序論

全体的に説明が簡潔すぎて、ボリューム感が足りない。エコトーンなど、研究の立ち位置を追加。GSに関しても増強。Intro と Prelim をきれいに書いて、方法までやってから書くべし。

1.1 研究背景

浅水域が存在量的にも重要だというふうにも見えるのですが、なぜこれまで測量が進んでこなかったのかをうまく説明できる流れにするのは難しそうなので、どちらかというと端っこでこれまで見逃されてきた、後回しにされてきたけれど、実は水と陸のエコトーンの部分にあたって生態的にも人の資源や観光利用において重要な場所であるというふうなことは書いていいのでしょうか。浅水域の定義みたいなものがあればいいと思いました。深さの範囲とか、川岸、湿地、砂浜、干潟などがそういう場所にあたると思います。浅水域の重要性をイメージできる画像もあれば。

エコトーンの文脈でより詳細に。まず、20intro の研究の意義でしっかり書いてから要約的にまとめる

地球表面の大部分を覆う水域、特に沿岸部や河川などの浅水域(Shallow Water)は、人間社会の経済活動、防災、生態系保全において極めて重要な役割を果たしている。河川管理における氾濫原の地形変状把握[Fujii2024_seigyu]、沿岸域管理[Pasquali2021_coastal-zone-management]、生態系の生息環境評価[Thomson2001_habitat-assessment]など、水底の詳細な地形データを観測する水深測量(Bathymetry)は非常に重要である。しかしながら、これらの領域は従来の測量技術では取得が困難な空白地帯となりがちであった。

ここは、20intro の関連研究で詳細に説明

伝統的な船舶搭載型マルチビームソナー(深浅測量・音響測深)は、一定の深度がある海域においては標準的な手法であるが、水深が極めて浅い河川や海岸線付近においては、船舶の座礁リスクなどの航行不可領域の存在により、その運用は著しく制限される。近年では、小型の無人水上艇(Unmanned Surface Vehicle: USV)を用いた水深測量が注目されているが[Giordano2015_sonar-bathymetry,Kurowski2019_survey-USV]、マルチビームの指向角の制限により、浅水域においては走査幅(Swath Width)が狭く、面的な測量を行うには時間的コストが高くなるという、非効率性の問題も生じる。一方で、航空機搭載レーザ測深(ALB: Airborne LiDAR Bathymetry)は[Saylam2018_ALB]、広域かつ高精度な計測が可能であるが、導入および運用コストが極めて高く、高頻度なモニタリングには不向きであるという経済的な障壁が存在する。

こうした背景の中、近年急速に普及したドローンなどの無人航空機(UAV: Unmanned Aerial Vehicle)を用いた写真測量(Photogrammetry)は、低コストかつ高解像度、高頻度なデータ取得が可能であることから、次世代の浅瀬測量技術として大きな期待を集めている。UAVにより空撮された多視点画像から、Structure-from-Motion(SfM)[schoenberger2016_colmap]、および Multi-View Stereo(MVS)

[Furukawa2010_PatchMVS·Furukawa2015_MVS]を用いて3次元形状を復元するアプローチは、陸部においては既に広い用途で実用化されている[Bemis2014_UAV-photogrammetry·Gomez2016_UAV-photogrammetry]。これらの技術を水中に適用する場合、その手法は空中からの水深写真測量(Photogrammetric Bathymetry)と呼ばれ、空中から撮像した多視点画像からの水中の三次元再構成問題と捉えることができる。水深写真測量には、光の反射や、水中での光の散乱・吸収による減衰、波による被写体の歪みなどの課題が存在するが、中でも最も根本的で重要な課題として取り組まれてきたのが光の屈折(Refraction)である。

1.2 研究課題

既存のSfM·MVSアルゴリズムの大部分は、幾何光学(Geometry Optics)を前提としている。すなわち、撮像(Image Sensing)のプロセスにおいて、観測対象となる被写体から発せられる光は、被写体からカメラ中心まで直進することを仮定する。しかし、UAVによる水中撮影においては、光は水中から空気中へ進む際に、水面と異なる媒質の境界でスネルの法則(Snell's Law)に従って屈折する。この物理現象によって、カメラから見た被写体の「見かけの位置」(Apparent Appearance)は、実際の位置よりも浅く、近く、歪ませる。屈折の影響を無視し、既存のSfM·MVSアルゴリズムを適用する場合、水深が実際よりも浅く評価される。これに対処するために、従来はSfM·MVSの出力結果に屈折率に基づく補正を適用する手法[westaway2001_PhotograBathy-multiply-n·Woodget2014_PhotograBathy-multiply-n]や、点群とカメラフレームの位置関係から推定する手法[Murase2008_refractiveCorrection·Dietrich2016_multi-angle-correction]、手動で計測した数カ所の真値をもとに出力結果の補正率を回帰で決定する手法

Yudhaさんなどの手法が回帰

が提案してきた。こうした手法は、屈折の影響を補正する一方で、屈折の物理的特性を完全にモデル化しているわけではなく、視線角度依存性や多視点間の整合性を厳密に扱えないため、幾何学的精度には限界があった。直近では、[Makris2024_refractive-aware-sfm]が、屈折の物理的特性を直接SfMの最適化に組み込むことで、より高精度な再構成を実現する手法を提案している。しかし、この手法では、疎な出力結果を補うために、USVなど高いコストを要する計測機器を用いた測量結果を用いた深層学習手法で補間する[Agrafiotis2019ISPRS_SVM-UAV-Bathymetry]必要があり、学習データ不足とフィールド依存性という問題がある。

さらに、近年のコンピュータビジョン(Computer Vision)・コンピュータグラフィックス(Computer Graphics)の領域では、Neural Radiance Fields(NeRF)[Mildenhall2021ECCV_NeRF]や3D Gaussian Splatting(3DGS)[Kerbl2023ToG_3DGS]といった、微分可能なレンダリング(Differentiable Rendering)を用いた新たな3次元表現・再構成手法が登場している。これらは任意視点における写真のようなリアルな新規視点合成(Novel View Synthesis:NVS)において卓越した性能を示し、照明依存性、時間軸方向へ拡張、幾何情報の抽出といった多種多様な課題を克服するよう、日進月歩の進化を遂げている

Survey論文でも引用しつく

。特に、NeRFのような陰的三次元表現(Implicit Representation)は計算コストが高く、幾何的な走査や解釈が容易ではない一方、3DGSは明示的な三次元Gaussian点群表現(Explicit Representation)を持ち、高速かつ直接的な形状操作が可能であるという利点を持つが、空気中から観測する水中屈折を

考慮した定式化は未だ十分になされていない。

1.3 研究の目的と貢献

GS の簡潔な説明。Explicit で高速。編集可能性や解釈可能度が高い。詳細は 42prelimGS で詳細に説明

これまでどんな問題に導入されてきたか、浅水域の写真測量の気体-液体屈折においてなぜ有効だと思われるか、イントロの中でも簡単に説明ができればいいと思いました。見る角度によって変化するものに対応とか、点を確率的なものに考えることで 0-1 の離散的ではなくより連続的な値が求められる?など。それが、結果としてはどうだったか、最後に考察で議論もしてもらいたいところです。

以上の背景から、本修士論文では、UAV 空撮画像からの水中 3 次元復元において、物理的な屈折モデルを Gaussian Splatting(GS) のパイプラインに直接統合することで、幾何学的正確性と写実的な外観再現を両立させる新たな枠組み (Refractive-Aware Gaussian Splatting) を提案・実証する。本手法の核心的な貢献は、水中の真の位置にある 3D Gaussian を、UAV 空撮画像中の見かけの位置へと解析的にマッピングするパラメータ変換にある。この変換は、微分可能であるたり、GS の最適化過程に直接組み込むことで、屈折を含む入力画像から直接、屈折のない三次元シーン (3D Scene) を推定し、屈折の影響を排した密な 3 次元形状と詳細なテクスチャ情報の復元を実現する。

日本語版の図を作成

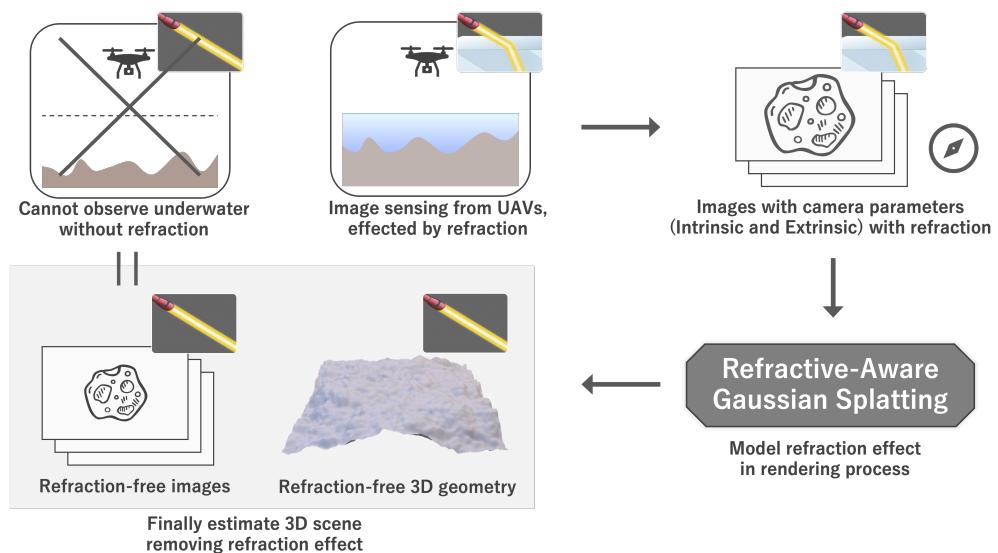


図 1-1: 空気中からの水深写真測量のタスクの概要。空中から水底を撮像したパラメータ既知の多視点画像を入力として、Refractive-Aware Gaussian Splatting を用いることで、屈折の影響を排除した水中の三次元シーンを再構成する。

検証においては、PBR レンダリング (Physically-Based Rendering:PBR) によってシミュレートされた合成データと実データの双方を用いて検証を行った。合成データでは、不観測地点からの再構成 3D モデルの視覚的品質を指す Peak Signal-to-Noise Ratio (PSNR) が 25 dB を超え、提案手法から抽出した幾何情報では真値との幾何的復元精度を表す F1 スコア (F1-score) で 94% を超える結果を確認した。(合成データは深度 10 m、撮影高度 10 m のスケールである。完全に平面の水面とカメラパラメータ既知を仮定し、屈折の影響のみを考慮し、反射や減衰、波など屈折以外の物理的影響は除外している。F1 スコア

は許容誤差 10 cm とした。) 実データにおいても、幾何的誤差 を確認するなど、実用的な空中からの水深写真測量としての方法を確認した。

1.4 論文構成

本論文は導入部である第 1 章を含め、全 6 章で構成される。第 2 章では、水深測量の概要と諸手法といった関連研究に関して述べる。第 3 章では、本研究の基盤となる Gaussian Splatting やコンピュータビジョンの諸手法の学術的背景と理論に関して記述する。第 4 章では、提案手法である Refractive-Aware Gaussian Splatting の詳細な理論と実装に関して記述する。第 5 章では、検証において用いた合成データと実データの双方を用いて、提案手法の性能を定性・定量的に検証する。

+ Field Work と実環境での検証、としてわけても良い

第 6 章では、研究の課題 (Limitation) と今後の課題 (Future Work) に関して述べ、研究の統括を行う。

第2章 研究背景と関連研究

子供に説明するような、だよ口調で論理の構造を作つてから、適切な文章に書き直す方式。書き直しは Gemini も有効に使う

基本的には、Goodnote に下書きしたのを写している。手書きしないと、スムーズに Output できない。

2.1 研究の意義

ecotone の概念を例に上げ、研究の適用先と意義を先に明確にする

Missing figure

ecotone の概念が分かるような参考の写真。陸から川底を取った写真(植生付き)、海岸線やサンゴ礁。(海の写真とかはどこから引用するんだろう)

2.2 水深測量の概要

研究の意義で説明した部分を削り、応用先に対比させるように記述

水深写真測量は重要だよ。地球の 7 割が水域だよ。にも関わらず、水底の詳細な地形データを観測する水深測量は大変で、海底のことは、宇宙や月のことよりも分からぬと言わがちだよ。近年 Seabed 2030 のような活動によって、地球の全域を測量しようとする気運が高まっているよ。

最近読んだ深海の地図をつくる? みたいな本を参照したいよ

それらはマルチビームソナーによる深浅測量によって行われるよ。これは音響測深とも呼ばれるよう、水中の音波を用いて水深を測量する手法であるよ。(一般的なセンシングで用いられる電磁波は、真空や薄い空气中では伝搬し情報を伝えられるが、水中では媒質が水であるため、その水中で減衰しにくい情報を使用するしかないのだ。的なことをかっこよく述べていよ) 昔は鎖を下ろして、航海という人類の活動の基盤の一つで、船の座礁などのリスクを管理する重要な海の地図だけど、それを作るのはたいへんやった。(歴史) 昔は、地点ごとに鎖を下ろして、深さを図るなど、一地点で半日かかりだった。(って本に書いてたけど、何を参照しよう)。その後、シングルソナーマルチビームにより、面的な測量が可能になった。マルチビームが最も一般的な Bathymetry の手法だよ。(Bathymetry もまたハイドロなんとかって分野に含まれる一つらしいが、うまく水に関する工学や自然科学の体系をわかっていないや) 近年では、小型の無人水上艇(Unmanned Surface Vehicle: USV) を用いた水深測量が注目されてい

るが [Giordano2015_sonar-bathymetry·Kurowski2019_survey-USV]、マルチビームの指向角の制限により、浅水域においては走査幅 (Swath Width) が狭く、面的な測量を行うには時間的コストが高くなるという、非効率性の問題も生じる。

Intro からのコピペなので、より詳しく説明する

一方、浅水域は、より人間社会に密接に影響を与えるよ。(人間は陸に住むしな) 洪水の影響を予測するためには、河川の形状データを用いて流体シミュレーションするし、伝統的河川治水方法(聖牛)を用いた河川マネジメントでは、増水のたびに変わる河川の地形を知りたいよ。(テンマの論文を引用! 具体的に想定するケースを明確に) また、生態系にはその生物の住む3次元的地形が密接に関連するよね。(こういう構造をシた場所だと産卵しやすいとか、外敵から身を守りやすいとか) あと、Geomorphology とかの学問でも重要でしょう。(体系的知識がないのでうまくまとめたい。Geology も) このように浅水域は海でも川でも、湖でも重要だよ。(特に共同の関連で私は川を念頭においているが、)

こうした浅水域を測量するのはとっても大変だよ。日本では、5年以内に一度、日本全国の川の水深測量を行う必要があるよ。人が実際にあって、棒尺で測定するのを本当にやるけど、危険で、空間解像度は低いし、アホほど Time Consuming だし、コストもばか。

浅水域では、深海とは異なり(深海ほど低くなくても)、電磁波(光)が完全に散乱、吸収されることなく、情報を残してくれるので、Remote Sensing のテクニックが使えるよ! [He2024_survey-shallow-bathymetry]
近年では、ALB が導入され、人に変わって測量できるよ。日本でも5年だか2年に1回だか図るよ(ここの河川マネジメントの資料は以前 Notion にまとめた。) LiDAR は Light Detection and Ranging の略で、光を用いて距離を測量する Passive な Remote Sensing だよ。水中で減衰しにくい緑とか青の波長域を照射し、水面で跳ね返るものと水中に潜って行くものを捉えることで、水底の形状を測量できるよ。深さもそこそこ行けるようになってきてるよ。密度も正確性も、申し分ないけど、航空機の飛行コストや機材の高価さによって、継続的な高周期のモニタリングは難しいよ。Passive な Remote Sensing として、他にも Satelite ベース(プラットフォーム)で、マルチスペクトルと、水中での波長の減衰率の差から測量する手法もあるけど、どうしても経験に頼ったり、Cite-Dependency が強かったり明示的なモデル化をもとに水中の3次元形状を推定することはできないよ。そこそこ深い場所を大規模に、は、有効だけど、より浅い地域を高解像度に、正確には、向かないんじゃないかな(ちゃんと論文を読んだことがないので、正確か分からぬ。しっかり厳しく、調査して教えてほしいな? Survey 論文にいろいろ書いていたよ。) Sar を使う手法もあるけど全然わからないや。

SAR も多少は

2.3 Photogrammetric Bathymetry

そこで、着目するのは、写真測量だよ。Computer Vision 技術の発達とドローンの普及によって正確な測量が誰でもできるようになったよ(商用ソフトウェア、OpenSource いっぱいあるね。近接写真測量の枠組みで話しているよ) UAV(+ RTKGPS) は、効率性、機材の安価さ、手軽な測量による高周期のモニタリングが可能だよ。

Close-Range Photogrammetry and 3D Imaging という、2020年くらいの大作が無料で Google Books 上で読めた。

https://books.google.co.jp/books?hl=en&lr=&id=L1DaEAAAQBAJ&oi=fnd&pg=PR5&dq=two+medium+photogrammetry&ots=7eT6ZRjJSx&sig=KJSIY_3Tdha35c5oIQRWRj1HVI&redir_esc=y#v=onepage&q&f=false

地上において、UAVによる写真測量は、広く実社会に普及しているよ。例えば、純粋な測量用途だけでなく、デジタルツインやメタバースといった3Dデータ活用の重要性が広く認識されつつある中で、災害状況把握、森林マネジメント、など様々な分野で利用されるよ [Bemis2014_UAV-photogrammetry Gomez2016_UAV-photog

水深測量に関しては、古い歴史があるよ。(1960年代などから、やってる人がいるよ。引用したいけどわからん)

既に書いてある内容で大丈夫だと思います。浅水域に対してこれまでの深浅測量の主流の応用が難しいこと、写真測量の可能性と、克服すべき点、あと限界などもうまく書ければ。音響と写真測量のメリットデメリットを表にできるといいかもしれません。

Missing figure

音響と写真測量のメリットデメリットの表

伝統的な船舶搭載型マルチビームソナー(深浅測量・音響測深)は、一定の深度がある海域においては標準的な手法であるが、水深が極めて浅い河川や海岸線付近においては、船舶の座礁リスクなどの航行不可領域の存在により、その運用は著しく制限される。近年では、小型の無人水上艇(Unmanned Surface Vehicle: USV)を用いた水深測量が注目されているが [Giordano2015_sonar-bathymetry Kurowski2019_survey-USV]、マルチビームの指向角の制限により、浅水域においては走査幅(Swath Width)が狭く、面的な測量を行うには時間的コストが高くなるという、非効率性の問題も生じる。一方で、航空機搭載レーザ測深(ALB: Airborne LiDAR Bathymetry)は [Saylam2018_ALB]、広域かつ高精度な計測が可能であるが、導入および運用コストが極めて高く、高頻度なモニタリングには不向きであるという経済的な障壁が存在する。

こうした背景の中、近年急速に普及したドローンなどの無人航空機(UAV: Unmanned Aerial Vehicle)を用いた写真測量(Photogrammetry)は、低コストかつ高解像度、高頻度なデータ取得が可能であることから、次世代の浅瀬測量技術として大きな期待を集めている。UAVにより空撮された多視点画像から、Structure-from-Motion (SfM) [schoenberger2016_colmap]、および Multi-View Stereo (MVS) [Furukawa2010_PatchMVS Furukawa2015_MVS] を用いて3次元形状を復元するアプローチは、陸部においては既に広い用途で実用化されている [Bemis2014_UAV-photogrammetry Gomez2016_UAV-photogrammetry]。これらの技術を水中に適用する場合、その手法は空中からの水深写真測量(Photogrammetric Bathymetry)と呼ばれ、空中から撮像した多視点画像からの水中の三次元再構成問題と捉えることができる。水深写真測量には、光の反射や、水中での光の散乱・吸収による減衰、波による被写体の歪みなどの課題が存在

するが、中でも最も根本的で重要な課題として取り組まれてきたのが光の屈折 (Refraction) である。

[Woodget2014_PhotograBathy-multiply-n] は、SfM ソフトの出力結果に単に「屈折率」を掛け合わせるというシンプルな補正を行うことで、ドローン (UAV) 画像を用いた水中写真測量の可能性を実証しました。これに対し [Dietrich2016_multi-angle-correction] は、見かけ上の点が視点によって異なること（視角依存性）を考慮し、画素ごとの光線の角度に基づいて 3 次元点を推定する「多角的な屈折補正」を提案しました。しかし、この手法はあくまで深さ（鉛直）方向のズレを補正するだけであり、水平方向の歪みについては無視されていました。そこで [Makris2024_refractive-aware-sfm] は、計算が終わった後に補正したり何度も計算を繰り返したりするのではなく、SfM の処理プロセス（パイプライン）そのものに屈折モデルを直接組み込むことに成功しました。この R-SfM は正確なカメラ位置推定と、疎な 3 次元点群を提供する。しかし、深層学習を用いた補間手法 [Alevizos2022_DL-shallow-bathymetry] を用いる必要があり、学習データ不足とフィールド依存性という問題がある。(もう少し Photogrammetric Bathymetry に関しては Survey するよ)

気体 - 液体屈折について。これも既に書いてある内容で大丈夫だと思います。また、ここからは方法でも詳しく説明する部分なので、イントロでどこまで書くか。浅水域の写真測量において克服すべき点の 1 つであること、屈折とはどういう現象か、これまで写真測量でどういう風に対処してきたか、測量を向上するために何が足りないか・何を考えるべきか

第3章 背景理論

3.1 画像に基づく3次元復元 (3D Reconstruction from Images)

画像を用いた測量は写真測量(Photogrammetry)と呼ばれ、古い歴史あり。<https://duplicate-3d.com/rd/2025-09-11-photogrammetry-history/> この記事を参考に。もともとは専門的な技能を持った人が、専用の機械(?)を用いて作成。

画像のデジタル化、イメージセンサの発明によりコンピュータビジョン(Computer Vision, CV)が始まる。Image Sensorから得た視覚情報から、人間や他の生物と同じように Computer や Machine に Scene を理解させる。

画像センサ(Image Sensor)が捉える情報は、本質的には3次元シーンから2次元平面への射影(Projection)である。この過程において、3次元空間の奥行き情報は1次元分欠落し、情報の「縮退」が発生する。3次元復元の主眼は、この失われた次元を幾何学的制約や事前知識(Priors)を用いて補完し、元のシーンの構造を逆問題として解くことにある。3次元情報は生物が自己が生きる世界を認識、理解するうえで必須の情報であり、ロボットの自己位置推定、環境認識などにも多大なニーズがあり、Computer Vision の主要タスクとして数多くの研究がなされ、今もめちゃくちゃレッドオーシャン。

古くから、単眼画像から形状を推定する手法として、輝度やテクスチャ、影などの手掛かりを利用する *Shape from X* ($X \in \{\text{shading, Silhouette, Texture, Focus, etc.}\}$) の研究が行われてきた。しかし、より頑健な復元を行うためには、視点移動を伴う複数枚の画像、あるいは動画(Image sequence)を用いて幾何的な整合性を元に手法が主流となった。この手法、タスクの総称を Structure from Motion (SfM) と呼ぶ。この SfM を起点とする、3D Vision for Geometry 手法の発達、高度な UI を備えた商用ソフト (Pix4D, Metashape, etc.) の出現、オープンソース化により、今日では Photogrammetry 技術は広く一般に普及した。

3.1.1 Structure from Motion

SfM は、複数の視点から撮影された画像群に基づき、カメラの内部・外部パラメータ(三次元的な動き)と、シーンの疎な(Sparse)3次元構造を同時に推定する手法である。Tomasi-Kanadeによる行列分解法に始まる。Hartley や Zisserman ら(2000年代)によって幾何学的理論の基礎が確立された。(VGGT のYoutube の対談動画で取り上げられていた。もう少し詳しくは、Harley 先生たちの本を参照。) 2016年に発表された COLMAP は、高い精度と汎用性、そして Open Source のプロジェクトとしての完成度から現在でもアカデミック分野でデファクトスタンダードとして広く利用されている。

COLMAP では、Pixel wise なんちゃらで、MVS の新規研究も実装されている

また、COLMAP により出力される、正確なカメラ内部パラメータ (Intrinsic) と外部パラメータ (Extrinsic) は、歪みのない画像 (Undistorted Images) は、後述する Dense Reconstruction や 新規視点合成タスク のための入力データとして、重要なパイプラインの一環を担っている。

In a typical incremental SfM pipeline, keypoints are firstly detected and matched across frames using feature detectors and descriptors such as SIFT [Lowe2004_SIFT]. Fundamental matrix F between two images is then calculated, commonly via eight-points algorithm [Hartley1997_8PointAlgorithm] combined with random sample consensus (RANSAC) [Fischler1987_RANSAC], which lead to camera pose recovery through singular value decomposition. New camera poses are iteratively registered via Perspective-n-Point algorithms, which align estimated 3D points with 2D features in new frames. Triangulation is subsequently applied to obtain additional 3D points from feature correspondences, and bundle adjustment (BA) [Triggs2000_BA] is finally performed to minimize reprojection error, refining both camera parameters and 3D points. 画像を貼るぞい!!!<http://theia-sfm.org/sfm.html>

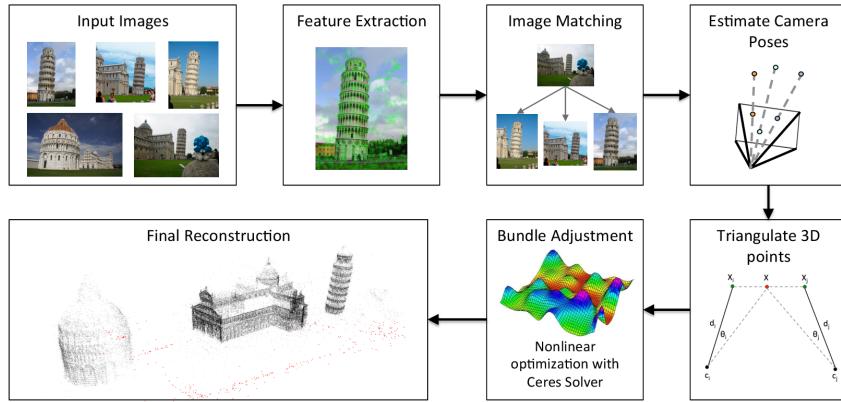


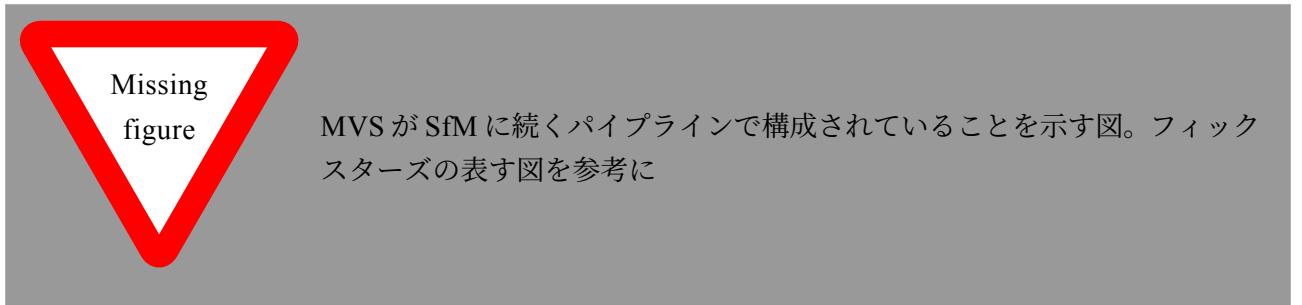
図 3-1: SfM Pipeline [fig:Theia-SfM]

3.1.2 Multi-View Stereo

SfM などによって得られたカメラパラメータが既知の多視点画像群から、各画素単位で密な深度推定により、高密度な三次元復元を行うのが Multi-View Stereo (MVS) である。SfM の出力三次元情報が疎な (Sparse) 点群であるのに対して、MVS は物体表面の密な (Dense) な三次元点群やメッシュ (Mesh) を推定することから、MVS は Dense 3D Reconstruction (密な三次元再構成) のタスクを行っていると言える。これを達成するため、各画素における深度 (Depth) や法線 (Normal) といった幾何パラメータを、画像間の整合性 (Photometric Consistency) に基づいて最適化する手法がとられる。本節では、パッチベース手法の古典である PMVS [Furukawa2010_PatchMVS] の概念を概観した後、現在のデファクトスタンダードなパイプライン (COLMAP [Schnberger2016ECCV_PatchMatchStereo]) の基礎となっている PatchMatch Stereo について詳述する。



図 3-2: PMVS Pipeline。[fig:Theia-SfM] から引用。左から右にかけてそれぞれの画像は、(1) 入力された多視点画像の一例; (2) 検出された特徴点; (3) 初期マッチング後のパッチ; (4) 拡張とフィルタリング後のパッチ; (5) メッシュモデル。



PMVS:

Patch-based Multi-View Stereo (PMVS)

[Furukawa2010_PMVS] によって提案された Patch-based Multi-View Stereo (PMVS) は、物体表面を微小な平面パッチの集合としてモデル化し、これを拡張・最適化することで密な 3 次元形状を復元する手法である。本手法は、特徴点マッチングによる疎な初期復元から出発し、信頼度の高いパッチを周囲に拡張(Expand)していくアプローチをとる、物体表面の形状を密に 3 次元推定するための先駆的かつ代表的な手法である。

cite では、現在の設定では名前が表示されずただ、[8] のように番号のみになるため、別の方や引用のフォーマットの設定を考える

PMVSにおける基本単位であるパッチ B は、単なる画像の矩形領域ではなく、対象物体の表面に接する微小平面（局所接平面）の近似として定義される。パッチ B は以下のパラメータを持つ。

- 中心座標 $p_B \in \mathbb{R}^3$: パッチの中心位置。
- 法線ベクトル $n_B \in \mathbb{R}^3$: パッチの向きを表す単位法線ベクトル。
- 参照画像 R_B : パッチ B を観測する画像の中で、光学的・幾何的に最も適した画像。
- 可視画像集合 V_B : パッチ B がオクルージョンなく観測可能であり、かつ相関スコアが閾値以上となる画像の集合。

PMVS の処理は、特徴点からの初期化(Initialization)の後、以下の 3 つのステップ、拡張(Expansion)、フィルタリング(Filtering)、最適化(Optimization)を反復することで行われる。

1. **Initialization (初期化)**: SfM と同様に、各画像から特徴点検出・マッチングを行う [Lowe2004_SIFT]。ここから得られる疎な点群を Seed Patch として、初期の法線とともに最初のパッチ群を生成するよ。前処理としてカメラパラメータの取得に SfM を用いている場合、これらの特徴点と、Triangulation(三角測量)によって得られる疎な三次元点群はそのまま使用できる。
2. **Expansion (拡張)**: 物体の表面が滑らかに連続しているという仮定に基づき、パッチをシーンの表面に沿うように増殖させていく。既存のパッチ B を参照画像 R_B および可視画像集合 V_B に投影し、その隣接画素に対応する空間領域にパッチが存在しない場合、新たなパッチ B' を生成する。この際、親パッチ B の法線 \mathbf{n}_B と深さ情報を初期値として継承させることで、テクスチャが弱い領域であっても、隣接する確度の高い領域から表面を「張り出して」いくことが可能となる。
3. **Filtering (フィルタリング)**: 拡張プロセスによって生じた誤ったパッチを除去する。以下の 3 つの基準が主に用いられる。
 - **Visibility Consistency**: 複数のパッチが同一の視線上に存在する場合、カメラに近い方を残し、隠蔽される奥のパッチを削除する。
 - **Photometric Consistency**: 正規化相互相關 (NCC) 等を用いた画像間の整合性スコアが一定以下のパッチを外れ値として破棄する。
 - **Number of Views**: パッチを安定して観測できるカメラの台数 $|V(p)|$ が最小閾値未満のものを信頼性不足として削除する。
4. **Optimization (最適化)**: 各パッチの位置 p_B と法線 \mathbf{n}_B を微修正し、画像間の整合性を最大化する。具体的には、パッチ B を可視画像 $I \in V_B$ へ投影して得られる画素値と、参照画像 R_B 上の画素値との間の Photometric Discrepancy を最小化するよう、非線形最適化を行う。

$$g(p) = \frac{1}{|V(B) \setminus R(B)|} \sum_{I \in V(B), I \neq R(B)} g(B, I, R(B))$$

ここで, $g(B, I_1, I_2)$ は、パッチ B に対する、画像 I_1 と I_2 の間の Photometric Discrepancy を測定するための関数であり、NCC などを使用することができる。

Photometric Discrepancy を日本語で説明すると、

この際、パッチの法線 $\mathbf{n}(p)$ を考慮して各画像をホモグラフィ変換することで、視点による透視歪みを補正し、より正確なマッチングを実現している。

以上のプロセスを収束するまで繰り返すことで、初期の疎な点群は徐々に密度を増し、最終的に物体表面全体を覆う密な点群が得られる。[Furukawa2010_PMVS] の手法は、大域的な最適化を行う PatchMatch Stereo 等と比較して局所的な貪欲法に近い性質を持ち、計算コストが高いが、拡張ステップによる表面の連続性利用が強力であり、密な復元を実現するための歴史的に重要な手法である。

3.1.3 PatchMatch Stereo

PMVS が、信頼できる「種 (Seed)」から局所的に表面を拡張していくアプローチであるのに対し、**[Bleyer2011BMVC_PatchMatchStereo]** が提案した PatchMatch Stereo は、画像上のすべての画素に対して、個別の 3 次元平面パラメータ（深度と法線）を推定する手法である。

従来の局所ステレオマッチング手法は、マッチングウィンドウ内の深度が一定である（カメラに対して平行な平面：Fronto-parallel window）と仮定することが一般的であった。しかし、この仮定は傾いた面や曲面において成立せず、再構成精度の低下や「階段状」のアーティファクトを生む原因となっていた。これに対し、Bleyer らは各画素のウィンドウを 3 次元空間内の傾いた平面（Slanted Support Window）としてモデル化した。しかし、各画素に対して最適な平面パラメータ（深度および法線の向き）を決定しようとすると、その探索空間は連続値であり無限大となるため、従来の総当たり的な探索や離散ラベルを用いる手法（Graph Cuts など）は適用できない。

この問題を解決するために導入されたのが、Barnes らによる最近傍探索アルゴリズム ”PatchMatch” の概念をステレオ視に応用した推論フレームワークである。

元論文をよりしっかりと読み込む。正しいか？

平面モデルとマッチングコスト: 左画像の各画素 Γ に対し、3 次元平面 f_Γ を割り当てる。平面 f_Γ は 3 つのパラメータ $(a_{f_\Gamma}, b_{f_\Gamma}, c_{f_\Gamma})$ を持ち、画素 Γ の座標 (Γ_x, Γ_y) における深度 d_Γ は以下の式で表される。

$$d_\Gamma = a_{f_\Gamma} \Gamma_x + b_{f_\Gamma} \Gamma_y + c_{f_\Gamma}$$

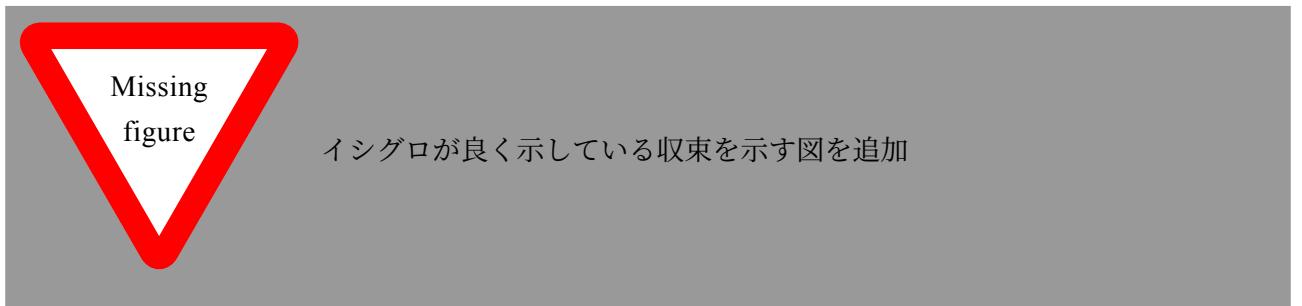
このモデルにより、画素ごとに異なる法線を持つ傾いた平面を表現でき、サブピクセル精度の深度推定が可能となる。最適化の目的は、各画素 Γ において、アグリゲーションコスト $m(\Gamma, f_\Gamma)$ を最小化する平面 f_Γ を、無限の候補空間 \mathcal{F} から見つけ出すことである。

$$f_\Gamma = \underset{f \in \mathcal{F}}{\operatorname{argmin}} m(\Gamma, f)$$

ランダム探索と伝播による推論: PatchMatch Stereo の核心は、ランダムな初期化状態から、空間的・視点的な相関を利用して「良い解」を画像全体に伝播 (Propagate) させるプロセスにある。アルゴリズムは、以下のステップを反復することで収束する。

1. **Random Initialization (ランダム初期化):** 初期状態では、すべての画素に対し、ランダムなパラメータを持つ平面（ランダムな深度と法線ベクトル）を割り当てる。
2. **Spatial Propagation (空間伝播):** 「隣接する画素は、同じ平面上に乗っている可能性が高い」という仮定を利用する。画像走査順序に従い、現在の画素 Γ とその近傍画素 Γ' （例えば左上の画素）を比較する。もし、近傍画素 Γ' の持つ平面 $f_{\Gamma'}$ を画素 Γ に適用した際のマッチングコストが、現在の平面 f_Γ よりも低くなるならば、画素 Γ は平面 $f_{\Gamma'}$ を自身の新たな推定値として採用し、コピーする。これにより、画像の一部で偶然「正解に近い平面」が見つかれば、その平面情報は波紋のように隣接画素へと広がり、領域全体が正しい平面で埋め尽くされていく。

3. **View Propagation (視点間伝播)**: ステレオ特有の拡張として、左右画像間の一貫性を利用する。左画像の画素 Γ の対応点である右画像の画素 Γ' が、より適切な平面パラメータを持っている場合、それを左画像座標系へ変換して取り込む。これにより、オクルージョン領域外での整合性が強力に担保される。
4. **Plane Refinement (平面の微調整)**: 伝播のみでは、既存の平面パラメータのコピーしか行われないため、真値へ到達できない。そこで、現在の平面パラメータに微小なランダム擾動 (Perturbation) を加え、コストが改善するかをテストする。反復が進むにつれて擾動の範囲を指数関数的に狭めていくことで、サブピクセルレベルでの高精度な収束を実現する。



この手法の利点は、巨大なコストボリューム (Cost Volume) をメモリ上に構築する必要がないため、高解像度画像や大きな視差範囲に対してもメモリ効率が良い点にある。また、連続空間での最適化を行うため、離散化に伴う量子化誤差が発生せず、極めて滑らかな曲面や急峻な傾斜面の復元に成功している。**Pixel wise View Selection** が、この手法をいかに MVS へ拡張したか説明する。

[Bleyer2011BMVC_PatchMatchStereo] は、基本的に 2 枚の画像 (ステレオペア) 間でのマッチングを前提としている。しかし、実際の SfM/MVS パイプラインでは、数十から数千枚の画像 (Multi-View) が入力され、かつそれらがインターネット上の写真のように撮影条件がバラバラな「非構造化 (Unstructured)」データである場合もある。

このとき、画像間でのパッチ選択は非効率!?!?

この課題に対し、[Schonberger2016ECCV_PatchMatchStereo] は、PatchMatch Stereo のフレームワークを多視点へ拡張する際、「画素ごとの視点選択 (Pixelwise View Selection)」という概念を導入することで解決を図った。

多視点ステレオにおける最大の課題は、ある参照画像の画素 Γ を復元するために、「どのソース画像を使うべきか」が画素ごとに異なる点である。画像全体で一律にソース画像を選んでしまうと、オクルージョンや解像度の不一致により、特定の画素ではマッチングが破綻してしまう。COLMAP では、以下の 3 つの幾何学的事前分布 (Geometric Priors) を確率モデルに組み込み、各画素 Γ が自分自身にとって最適なソース画像を動的に選択しながら推論を行う。

1. **Triangulation Prior**: 十分なベースラインを持ち、三角測量の精度が保証される角度 (Triangulation Angle) で撮影された画像を優先する。角度が小さすぎる (視点が近すぎる) 画像は深度推定の不確定性が高いため除外される。

2. **Resolution Prior**: 参照画像とソース画像で、対象を捉えている解像度（画素密度）が類似している画像を優先する。極端に解像度が異なる画像間でのマッチングは、エイリアシング等の問題を引き起こすためである。

3. **Incident Prior**: 推定された法線ベクトルに対し、カメラ視線が正対に近い（斜めすぎない）画像を優先する。これにより、極端な浅い角度から撮影された信頼性の低い画像の影響を排除する。

Schönberger らは、これらの幾何学的尺度とフォトメトリックな整合性を統合した確率的グラフィカルモデルを構築し、PatchMatch の反復プロセスの中で「深度・法線の推定」と「最適なソース画像の選択」を同時に最適化する手法を確立した。これにより、COLMAP は極めてノイズの多い非構造化データセットに対しても、Robust な密な三次元再構成を可能にした。

一方、Geometric Optics による光の直進性を仮定するが、これは屈折のある Scene では成立しない。(Reflection も一般的に苦手)

3.1.4 Feed Forward 3D Reconstruction

<https://gemini.google.com/app/d37c008e08350238>

DINO などの深層学習ベースの特徴抽出を解説。自分の提案手法の前処理にも、水面マスクに Dino を使用する

本研究の本筋である最適化ベース (Optimization-based) の手法とは対照的に、近年 (2023-2025 年)、大規模データセットと Deep Learning、特に Transformer アーキテクチャの発展により、幾何学的計算を推論 (Inference) として解く「フィードフォワード型 (Feed-Forward)」の手法が急速に台頭している。従来の手法 (SfM/MVS や NeRF/3DGS) が、入力シーンごとにパラメータを反復的に更新して解を探索するのに対し、フィードフォワード型の手法は、学習済みの膨大な事前知識 (Priors) を用いて、単一の順伝播処理のみで 3 次元構造を回帰するデータ駆動型 (Data-driven) のアプローチである。

このアプローチの代表例として、以下の手法が挙げられる。

- **Depth Anything [yang2024depthanything]**: 6200 万枚以上の画像から学習された基盤モデル (Foundation Model) であり、DINOv2[<empty citation>] バックボーンを活用することで、テクスチャのない領域や未知のシーンに対しても極めてロバストな单眼深度推定を実現している。しかし、出力はスケール不確定性を伴う 2.5 次元表現に留まり、多視点間での厳密な幾何学的整合性は保証されない。
- **DUST3R / MASt3R [wang2024dust3rleroy2025mast3r]**: 従来の SfM パイプライン (特徴点抽出、マッチング、バンドル調整) を完全に排除し、2 枚の画像から直接「ポイントマップ (Point Map)」を回帰する手法である。これにより、カメラパラメータを事前に与えることなく (Unposed)、エンドツーエンドでの 3 次元形状および Point Map からのカメラ姿勢の逆推定が可能となった。特に MASt3R は、。

MASt3R ってそんな手法だっけ?

- **VGGT (Visual Geometry Grounded Transformer) [wang2025vggt]**: CVPR 2025 にて Best Paper Award を受賞した、現時点での到達点といえる手法である。VGGT は、任意の枚数の画像を入力とし、カメラパラメータ、深度、点群、そして追跡情報（Tracks）の全てを同時に推論する。

これらの手法は、計算速度とロバスト性において革新的であるが、本研究が扱う「物理的に正確な表面再構成」の観点からは明確な限界も存在する。フィードフォワード手法はその性質上、学習データに深く性能を依存する。

そのため、屈折（Refraction）や透明物体（Transparency）を含むシーンにおいて、これらの手法は破綻しやすい。学習データ（ScanNet 等）に透明物体が十分に、かつ物理的に正確なアノテーションと共に含まれていないため、フィードフォワードモデルはガラス表面を背景と混同したり、深度を平滑化してしまう傾向がある。

根拠薄し

対して、本研究で用いる Gaussian Splatting 等の最適化ベース手法は、屈折率（IOR）やスネルの法則を明示的にモデル化することで、こうした物理現象を正確に逆算することが可能である。

しかし、動画生成 AI が流体や光の反射といった物理法則をデータから獲得しつつある現状 [quan2025transparent] を鑑みると、将来的にはフィードフォワード手法も十分なデータスケールによって屈折を「学習」する可能性は否定できない。現時点では、幾何学的整合性と物理的忠実性を担保するためには、依然として物理モデルに基づく最適化が不可欠である。



Missing figure

Optimization-based 手法と Feed-Forward 手法の処理フロー比較図：Iterative なループを持つ前者と、Single Pass で完結する後者の対比。また、透明物体に対する挙動の違い（透過してしまうか、屈折を考慮するか）の概念図。

3.2 Inverse Rendering as 3D Reconstruction

Inverse Rendering

<https://gemini.google.com/app/0bb4c009393122eb>

コンピュータグラフィックス (CG) の古典的な課題は、3 次元のシーン記述（形状、材質、光源）から 2 次元の画像を生成することである。この過程は「Forward Rendering（順方向レンダリング）」と呼ばれ、一般的にはレンダリングと呼ぶ。これを数学的な関数 R と見なすと、シーンパラメータ Θ から画像 I への写像 $I = R(\Theta)$ として表現できる。これに対し、Inverse Rendering（逆レンダリング）は、観測された画像 I_{obs} から、それを生成したシーンパラメータ Θ を推定する逆問題 $\Theta = R^{-1}(I_{obs})$ を解くことである。観測された画像群から 3D シーンを推定するという意味で、Inverse Rendering は画像からの三次元再構成の一例であるといえる。

しかし、物理的な世界から画像への射影は情報の損失（奥行きの消失、法線とテクスチャの混同など）

を伴うため、この逆問題は本質的に不良設定問題（Ill-posed problem）となる。この困難な問題を、最適化手法を用いて解くための強力なフレームワークとして登場したのが、Differentiable Rendering（微分可能レンダリング）である。

Missing figure

Rendering と Inverse Rendering を表す図

3.2.1 Analysis-by-Synthesis and Novel View Synthesis

現代の Inverse Rendering の多くは、Analysis-by-Synthesis というアプローチを採用している。これは、パラメータ Θ を直接回帰するのではなく、レンダリングされた画像 $R(\Theta)$ と観測画像 I_{obs} との間の再構成誤差 \mathcal{L} を最小化する最適化問題として定式化される。

$$\Theta^* = \underset{\Theta}{\operatorname{argmin}} \mathcal{L}(R(\Theta), I_{obs})$$

ここで \mathcal{L} は損失関数 (L2 ノルムや Perceptual Loss) である。この最小化問題を勾配降下法 (Gradient Descent) で解くためには、レンダリング関数 R がパラメータ Θ に関して微分可能である必要がある。すなわち、3D シーンパラメータ Θ に関する損失関数 \mathcal{L} の勾配 $\frac{\partial \mathcal{L}}{\partial \Theta}$ を計算できなければならない。連鎖律 (Chain Rule) を適用すると、勾配は以下のように分解される。

$$\frac{\partial \mathcal{L}}{\partial \Theta} = \frac{\partial \mathcal{L}}{\partial I} \cdot \frac{\partial I}{\partial \Theta}$$

前半の $\frac{\partial \mathcal{L}}{\partial I}$ (画像の画素値に対する Loss の勾配) は容易に計算できるが、後半の $\frac{\partial I}{\partial \Theta}$ (シーンパラメータの変化が画素値にどう影響するか) の計算は一般的な三次元形状表現である Mesh では難しい。
いらないと思う

??の定式化によって、

微分可能レンダリングによる最適化

逆レンダリングを勾配法 (Gradient Descent) によって解くためには、レンダリングプロセスが微分可能 (Differentiable) である必要がある。(なんで勾配法で解く必要があるのか土木人に分かるよう説明してあげないと。DL の発達とその最適化手法の根底にあることを言及) 微分可能レンダリング (Differentiable Rendering, DR) の枠組みでは、以下のループによって 3 次元シーンが最適化される。

1. **レンダリング:** 現在推定されている 3 次元シーン (形状・外観) から、仮想的な視点で画像を生成する。

2. **損失計算:** 生成された画像 (Rendered Image) と、実際に撮影された正解画像 (Source Image) との間の誤差 (Loss) を計算する。
3. **誤差逆伝播:** 微分可能なレンダラを介して、誤差の勾配 (Gradient) を 3 次元シーンのパラメータへと連鎖律に従って伝播させる。
4. **更新:** 勾配を用いて、3 次元シーンのパラメータを逐次更新する。

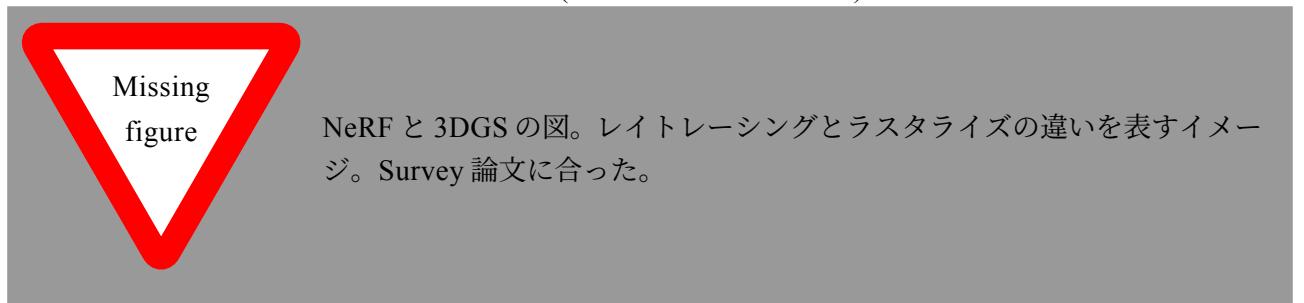
この手法は、幾何学的な特徴点のみならず、画像内の全画素の情報を最適化に利用できる従来の Bundle Adjustment も 2 次元画像座標上の特徴点を介して、3 次元シーンを構成する点群を再投影誤差を介し最小化するという点で、微分可能レンダリングの一種と捉えることができる。

Novel View Synthesis

従来の SfM や MVS の手法は Geometric Reconstruction として、主に幾何情報の復元に焦点を当てていた。実際の Scene は幾何情報 (Geometry) に加え、照明 (illumination)、テクスチャ (texture)、BRDF などでモデル化される表面特性などを含む。新規視点合成というタスクでは、ポーズを付与した画像から、その視点外からの新視点の画像を推定するタスクである。SfM や MVS が Geometric Reconstruction とするならば、新規視点合成は Appearance Reconstruction となる。Appearance には当然、Geometry の情報も含まれているため、Appearance Reconstruction は Geometry Reconstruction のタスクを暗に含んでいると言える。図を挿入して上げるぞ。(CG のレンダリングプロセスと、ベン図)

Neural Radiance Fields (NeRF)

微分可能レンダリングの代表例が NeRF (Neural Radiance Fields) である。



NeRF では、シーンをボクセルやポリゴンといった明示的な幾何構造ではなく、多層ペーセプトロン (MLP) を用いた「連続的な輝度場 (Radiance Field)」として表現する。これは、点群やメッシュといった、シーンの幾何情報を直接的にエンコードする Explicit Representation ではなく、関数を介して表現する Implicit Representation の一種である。



Missing
figure

NeRF の MLP の構造も図で載せたい。引用できる

具体的には、空間上の座標 (x, y, z) と視線方向 (θ, ϕ) を入力とし、その点における放射輝度（RGB）と体積密度 (σ) を出力する関数 f_Θ を学習する。この表現を用いて、以下の Volumetric Rendering によりピクセル値を算出する。

$$C(\mathbf{r}) = \int_{t_n}^{t_f} T(t) \sigma(\mathbf{r}(t)) \mathbf{c}(\mathbf{r}(t), \mathbf{d}) dt$$

ここで $T(t)$ は透過率、 $\mathbf{r}(t)$ はレイ上の点を表す。この積分プロセスは離散的なサンプリングによって近似され、微分可能な形で実装される。ボリュメトリックレンダリングは、元来、煙や炎などの不均等な媒質を通過する光の伝播をモデル化を可視化するために開発されたものだが、メッシュなどのラスタライズに含まれる離散性がなく、微分可能性を保つとともに、汎用的な Scene 表現に用いられる。これにより、複雑な光学的特性（反射や半透明など）を含むシーンにおいても、エンドツーエンドでの高精度な再構成が可能となった。

ピクセル色の決定には、古典的なボリュームレンダリングの原理が応用される。仮想的なカメラから放たれたレイ（光線）に沿って空間をサンプリングし、各サンプル点での密度 σ を重みとした放射輝度 \mathbf{c} の積分値を算出することで、最終的なピクセル値が決定される。この積分計算は離散的な総和として近似されるが、演算過程の全てが微分可能に保たれているため、レンダリング画像と実画像の二乗誤差を損失関数とし、MLP の重みをエンドツーエンドで最適化できる。しかし、NeRF は写真のような忠実度の自由視点合成を実現した一方で、実用上の重大な障壁も露呈させた。1 ピクセルの描画のためにレイに沿った数百回の MLP 推論を要する計算負荷の高さは、リアルタイムレンダリングを困難にし、またシーンごとに日単位の学習を要する点は、大規模なデータセットへの適用を制限している。また、3 次元シーンが MLP にエンコードされる点で、その解釈や編集可能性が乏しく、測量用途としての適用性に欠ける。これらの課題、すなわち「計算資源の集約性」と「暗黙的表現による編集の困難性」を克服しようとする動機が、後の明示的な幾何プリミティブを用いる手法への回帰を促すこととなった。

3.3 Gaussian Splatting

3D Gaussian Splatting

Missing figure

3DGS の図。3DGS のパイプラインの図

Missing figure

NeRF と 3DGS の分類図。レンダリング速度と、データサイズの 2D プロット。Barron さんのレクチャー動画から。学習速度も入れれば、2 次元に収まらない。。2 個いるかな？できれば、MVS とともに含めて示せねばいい。それだったら表が美しい

3D Gaussian Splatting (3DGS) は、新規視点合成において最先端 (SOTA) の結果を達成している近年の手法である [Kerbl2023ToG_3DGS]。その特徴は、フォトリアリスティックかつ高忠実度な 3 次元シーンのキャプチャ能力、高速な学習時間、そしてリアルタイムレンダリングにある。明示的な (Explicit) 3 次元表現である 3DGS は、Visual-SLAM [Yan2024CVPR_GS-SLAM·Zheng2025CVPR_WildGS-SLAM·Matsuki2024CVPR_GS-SLAM]、アバター生成 [Moreau2024CVPR_HumanGaussianSplatting·Shao2024CVPR_GaussianAvatar]、フィードフォワード型 3 次元再構成 [Chen2024ECCV_MVSplatting] など、幅広いタスクへの適用に成功している。その可能性はさらに広がり、衛星画像からの数値表層モデル (DSM) 生成 [Aira2025CVPR_EOGS]、自動運転、そして水中 3 次元再構成 [li20243DV_watersplatting] といった様々な実世界アプリケーションへと拡張されている。

Missing figure

GS が実際に椎円体のプリミティブで構成されていることを示す図。これは自分の発表で使用した寮の中庭の画像でいいかな

3DGS のパイプラインは主に、レンダリングを行うフォワードパスと、最適化を行うバックワードパスの 2 つの段階で構成される。フォワードパスでは、3 次元ガウス分布 (3D Gaussians) の集合をラスタライズして画像を合成する。各ガウス分布は、中心位置 $\mathbf{p} \in \mathbb{R}^3$ 、不透明度 $\alpha \in [0, 1]$ 、球面調和関数 (SH) によって表現される視点依存の色係数 $\mathbf{c}(\mathbf{p}, \mathbf{t}_i) \in \mathbb{R}^3$ 、および 3 次元共分散行列 $\Sigma^{3D} \in \mathbb{R}^{3 \times 3}$ という、最適化可能なパラメータセットによって定義される。共分散行列 Σ^{3D} は、スケーリングベクトル $\mathbf{s} \in \mathbb{R}^3$ から構成されるスケーリング対角行列 $\mathbf{S} \in \mathbb{R}^{3 \times 3}$ と、回転クォータニオン (回転行列 $\mathbf{R} \in SO(3)$) として

表現)を用いて以下のように構成される：

$$\Sigma^{3D} = \mathbf{R} \mathbf{S} \mathbf{S}^\top \mathbf{R}^\top$$

点 $\mathbf{x} \in \mathbb{R}^3$ に対する対応する非正規化ガウス分布関数は以下で与えられる：

$$G(\mathbf{x}) = \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{p})^T (\Sigma^{3D})^{-1} (\mathbf{x} - \mathbf{p})\right)$$

フォワードパスにおいて、あるカメラ視点からの画像をレンダリングするために、これらのガウス分布はまず外部パラメータ行列 $[\mathbf{W}|t]$ を用いてワールド座標系からカメラ座標系へと変換される。ここで、 $\mathbf{W}_{\text{view}} \in \mathbb{R}^{3 \times 3}$ は視点回転行列、 $t \in \mathbb{R}^3$ は平行移動ベクトルである。ガウス分布の中心位置 \mathbf{p} と 3 次元共分散行列は以下のように更新される：

$$\begin{aligned}\mathbf{p}_{\text{cam}} &= \mathbf{W}\mathbf{p} + t \\ \Sigma_{\text{cam}}^{3D} &= \mathbf{W}\Sigma^{3D}\mathbf{W}^\top\end{aligned}$$

[Zwicker2001_EWA-volume-splatting] の提案した射影手法に従い、カメラ空間における 3 次元共分散行列 Σ_{cam}^{3D} は 2 次元画像平面へと射影される。これは透視投影の一次近似(アフィン近似)のヤコビ行列 \mathbf{J} を用いて行われ、2 次元共分散行列 Σ^{2D} が得られる：

$$\Sigma^{2D} = \mathbf{J}\Sigma_{\text{cam}}^{3D}\mathbf{J}^\top$$

各ピクセルの最終的な RGB 値 $\Gamma \in \mathbb{R}^3$ は、射影されたガウス分布をアルファブレンディングすることでレンダリングされる。ピクセルと重なるガウス分布の集合は、まず深度に基づいて手前から奥へとソートされ、視点依存色が以下のように累積される：

$$\begin{aligned}\Gamma(\mathbf{x}) &= \sum_{k=1}^K \mathbf{c}_k \alpha_k^{\text{pixel}} \prod_{j=1}^{k-1} (1 - \alpha_j^{\text{pixel}}) \\ \text{where } \alpha_k^{\text{pixel}} &= \alpha_k G_k^{2D}\end{aligned}$$

ここで、 k はピクセルに重なる整列されたガウス分布の集合のインデックスである。

バックワードパスでは、最適化により測光誤差(Photometric loss)を最小化する。これは $\mathcal{L}_1(\Gamma, \Gamma_{gt})$ 損失と D-SSIM 損失 $\mathcal{L}_{\text{D-SSIM}}(\Gamma, \Gamma_{gt})$ [Zhou2004_SSIM] の加重和である：

$$\mathcal{L} = (1 - \lambda)\mathcal{L}_1 + \lambda\mathcal{L}_{\text{D-SSIM}}$$

したがって、最適化問題は以下のように定式化される：

$$\underset{p, R, s, c, \alpha}{\operatorname{argmin}} \quad \mathcal{L} = \mathcal{L}(\Gamma, \Gamma_{gt})$$

これらの定式化により、パイプライン全体が完全微分可能となり、パラメータ $\Theta = \{\mathbf{p}, \mathbf{R}, \mathbf{s}, \mathbf{c}, \alpha\}$ は勾配降下法によって最適化可能となる。その最適化に要する学習時間は、30k のイテレーションによって 1 時間以内となり、当時 NVS の Sota であった Mip-NeRF360 [Barron2022CVPR_Mip-NeRF360] に比較して 10 倍以上の高速化を達成した。加えて、Ray-Tracing に比較し、既存の GPU 描画パイプラインの性能を引き出す Tile-Based レンダリングによって、100 fps 以上のリアルタイムレンダリングを実現したことで、インタラクティブな Scene の可視化が可能となった。このプロセスを通じて得られた 3 次元ガウス分布の集合は、3 次元シーンを高忠実度で捉えることができる。

しかし、このパイプライン全体はピンホールカメラモデルと透視投影に依存しており、光が直線的に進むことを根本的な前提としている。この前提は、空気と水の境界での屈折が深刻な幾何学的矛盾を引き起こし、再構成の失敗につながるような、複数の媒質が介在する環境においては成立しない。

この制限にもかかわらず、3DGS は NeRF のような陰的表現 (Implicit representations) と比較して、この課題に対処するのに独自に適している。これは、ガウスプリミティブの明示的な性質が、直接的な物理モデリングに対して非常に開放的であるためである。これにより、屈折の法則を数学的に定式化し、シーン表現の幾何学的パラメータそのものに直接適用することが可能となる。

3.3.1 2D Gaussian Splatting

不透明度 α_i の計算方法が 3DGS とは異なる点です。2DGS では、レイとプリミティブ k の平面との交点 \mathbf{p}_{inter} を求め、その交点の「ガウス中心からのローカル座標系での距離」に基づいて評価します。

2DGS において、シーンは配向された平らな 2 次元ガウス分布 (Surfel) の集合として表現される。ピクセル座標 (u, v) を通過する視線 (Ray) $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ は、深度順にソートされた N 個のガウス平面と交差し、その交点における評価値が画像形成に寄与する。

3.4 戻り値の数理的解析

3.4.1 レンダリングされた色 (render_colors)

型: $[..., C, H, W, 3]$ (RGB)

これはボリュームレンダリング方程式の離散化表現である。色は以下の Front-to-back アルファブレンディングにより計算される。

$$C(\mathbf{r}) = \sum_{i \in \mathcal{N}} c_i \alpha_i T_i, \quad T_i = \prod_{j=1}^{i-1} (1 - \alpha_j)$$

ここで、透過率 T_i はレイが i 番目のガウス分布に到達する確率を表す。2DGS の特異性は不透明度 α_i の導出にある。レイと第 k プリミティブ平面との交点 \mathbf{p}_{inter} をローカル座標系へ変換した変位ベクトルを \mathbf{u} とすると、 α_i は以下のように定義される。

$$\alpha_i = o_i \exp \left(-\frac{1}{2} \mathbf{u}^\top \Sigma_{2D}^{-1} \mathbf{u} \right)$$

ここで、 o_i は学習可能な不透明度パラメータ、 Σ_{2D} は 2 次元スケーリングと回転行列から構成される局所的な分散行列である。

3.4.2 蓄積アルファ (render_alpha)

型: [..., C, H, W, 1]

レイに沿った総不透明度であり、物理的にはレイが物体によって遮蔽された確率を表す。これは最終的な透過率 T_{final} の補数となる。

$$A(\mathbf{r}) = \sum_{i \in \mathcal{N}} \alpha_i T_i = 1 - T_{final} = 1 - \prod_{i=1}^N (1 - \alpha_i)$$

3.4.3 ボリューム法線 (render_normals)

型: [..., C, H, W, 3]

2DG が「幾何学的に正確」であるための核心的な項である。各 2 次元ガウス円盤の固有法線 \mathbf{n}_i を、色情報と同様の重みでブレンディングしたものである。

$$\mathbf{n}_{vol}(\mathbf{r}) = \sum_{i \in \mathcal{N}} (\mathbf{R}_i \cdot \mathbf{z}_{local}) \alpha_i T_i$$

ここで \mathbf{R}_i はクオータニオンから導出される回転行列、 $\mathbf{z}_{local} = (0, 0, 1)^\top$ はローカル座標系における法線ベクトルである。この \mathbf{n}_{vol} は、後述する表面法線との整合性を取る正則化 (Normal Consistency Loss) に使用される。

3.4.4 表面法線 (surf_normals)

型: [..., C, H, W, 3]

レンダリングされた深度マップの幾何学的勾配から算出される法線であり、「見かけ上の形状」を表す。まず、期待値深度 (Expected Depth) $D(u, v)$ を計算する。

$$D(u, v) = \sum_{i \in \mathcal{N}} t_i \alpha_i T_i$$

ここで t_i はカメラ原点から交点までの距離である。 \mathbf{n}_{surf} は画像空間における勾配のクロス積として近似される。

$$\mathbf{n}_{surf} \propto \frac{\partial D}{\partial u} \times \frac{\partial D}{\partial v}$$

3.4.5 歪み (render_distort)

型: [..., C, H, W, 1]

Mip-NeRF 360 で提案された Distortion Loss の L1 変種である。レイ上の密度分布が局所的に集中(=明確な表面が存在)することを奨励する。

$$\mathcal{L}_{dist}(\mathbf{r}) = \sum_{i,j} w_i w_j |t_i - t_j| + \frac{1}{3} \sum_i w_i^2 s_i$$

(ただし実装上は計算効率化のため、上記式の簡易版が用いられる場合がある)

3.4.6 中央値深度 (`render_median`)

型: $[..., C, H, W, 1]$

累積透過率が閾値 0.5 に達した地点の深度。期待値深度におけるエッジ付近のアーティファクト (Flying Pixels) を抑制し、TSDF Fusion 等によるメッシュ再構成において堅牢な結果を提供する。

$$t_{median} = \operatorname{argmin}_t \left| \int_0^t \alpha(s) T(s) ds - 0.5 \right|$$

3.5 実装上の設計指針 (Engineering Advice)

Google Research の実装経験に基づき、安定したパイプライン構築のための重要な指針を以下に示す。

3.5.1 Ray-Plane Intersection の特異点処理

視線ベクトル \mathbf{d} と平面法線 \mathbf{n}_i が直交に近い場合 (Grazing Angle)、交点計算は数値的に不安定となる。CUDA カーネル内では、内積値に対するハードクリッピングを推奨する。

if $|\mathbf{d} \cdot \mathbf{n}_i| < \epsilon$ (e.g., $\epsilon = 0.05$), discard intersection.

3.5.2 幾何再構成のための深度モード

メッシュ抽出を目的とする場合、`depth_mode` の選択は重要である。浮遊ノイズの少ない高品質なメッシュを得るためにには、`render_median` の出力を信頼するか、あるいは正則化項を含んだ RGB+ED モードでの学習結果を用いることが望ましい。また、微分可能レンダリングにおいて `surf_normals` を利用する際は、計算グラフの切断 (`detach`) を適切に管理する必要がある。